BAYESIAN OPTIMIZATION ASSISTED UNSUPERVISED LEARNING FOR EFFICIENT INTRA-TUMOR PARTITIONING IN MRI AND SURVIVAL PREDICTION FOR GLIOBLASTOMA PATIENTS

Yifan Li, Chao Li, Stephen Price, Carola-Bibiane Schönlieb, Xi Chen

xc841@bath.ac.uk

ABSTRACT

Glioblastoma is profoundly heterogeneous in microstructure and vasculature, which may lead to tumor regional diversity and distinct treatment response. Although successful in tumor sub-region segmentation and survival prediction, radiomics based on machine learning algorithms, is challenged by its robustness, due to the vague intermediate process and track changes. Also, the weak interpretability of the model poses challenges to clinical application. Here we proposed a machine learning framework to semi-automatically fine-tune the clustering algorithms and quantitatively identify stable sub-regions for reliable clinical survival prediction. Hyper-parameters are automatically determined by the global minimum of the trained Gaussian Process (GP) surrogate model through Bayesian optimization(BO) to alleviate the difficulty of tuning parameters for clinical researchers. To enhance the interpretability of the survival prediction model, we incorporated the prior knowledge of intra-tumoral heterogeneity, by segmenting tumor sub-regions and extracting sub-regional features. The results demonstrated that the global minimum of the trained GP surrogate can be used as sub-optimal hyper-parameter solutions for efficient. The sub-regions segmented based on physiological MRI can be applied to predict patient survival, which could enhance the clinical interpretability for the machine learning model.

***Index Terms*—** Tumor sub-region imaging, Bayesian optimization, unsupervised learning, MRI, Glioblastoma

1. INTRODUCTION

Glioblastoma is a highly aggressive brain cancer characterized by the complexity of tissue microstructure and vasculature within tumor. Previous research shows that multiple intra-tumoral sub-regions exist and have distinct treatment response. Therefore, this intra-tumoral heterogeneity, leading to discrepancy in tumor composition among patients, poses significant challenges to patient individualized treatment and outcome prediction.[1].

SP acknowledges NIHR Career Development Fellowship (CDF-18-11-ST2-003). CBS acknowledges EPSRC (EP/M00483X/1 and EP/N014588/1); CL acknowledges CRUK biomarker grant (CRUK/A19732).

Magnetic resonance imaging (MRI) is the standard tool for disease management. As an emerging technique, radiomics could extract useful features from MRI for diagnosis and outcome prediction using machine learning (ML) techniques[2]. However, the robustness of radiomics is challenged by the stability of ML algorithms. Further, the clinical interpretability of radiomics needs to be improved for clinical translation.

In this paper, we proposed a semi-automatic ML pipeline to identify tumor sub-regions and predict patient survival using physiological MRI, i.e., perfusion (pMRI) and diffusion MRI (dMRI), which provide quantitative measures regarding the different perspectives of tumor biology. Potential drawbacks of the implemented unsupervised clustering, such as instability and unreliability, were addressed by incorporating a mathematical stability measure and an auto-encoder neural network. Moreover, a Bayesian optimization framework was also introduced to both efficiently speed up the fine-tuning process and improve the robustness of the clustering algorithm. To further enhance the interpretability, we designed a clinically sensible feature set to measure the inter-relation among the co-existing sub-regions for survival prediction.

2. PROBLEM FORMULATION

We used below quantitative maps calculated from the pMRI and dMRI that have been co-registered to the post-contrast T1-weighted images, with $\mathbf{v}$ describes the 3-D coordinate (v_x, v_y, v_z) of each voxel; $\mathbf{r}$ represents the relative cerebral blood volume (rCBV), calculated from pMRI; and $\mathbf{p}$, $\mathbf{q}$ denotes the isotropic and anisotropic component of dMRI respectively. The MRI of the ith patient can be denoted as $\mathcal{D}_i = (\mathbf{v}, \mathbf{r}, \mathbf{p}, \mathbf{q})$. Additionally, patient survival t_i can also be collected along with $\mathcal{D}_i$.

Given the training data $\mathcal{D}_i$ and its corresponding machine learning (ML) 'feature' t_i, one may apply various ML algorithms to train the prediction model. However, this can be problematic from the clinical perspective due to the model's unstable and unreliable black-box features, e.g., small changes of the training data may result in significant differences in both the models and prediction.

2.1. Sub-region segmentation via unsupervised learning

Unsupervised clustering algorithms have been successfully applied to segment clinically interpretable tumor sub-regions [3]. In this paper, partition-based clustering K-means and model-based clustering Gaussian mixture model (GMM) [4] were adopted for sub-region clustering. Instead of fixing K-means and GMM hyper-parameters (e.g. cluster number N_c), we treated them as unknown variables $\boldsymbol{\theta}$ to be optimized. The goal was to cluster data set $\mathcal{D}$ over patients and obtain N_c clusters (sub-regions) $\{C_1, \cdots C_m, \cdots C_{N_c}\}$.

2.1.1. Cluster stability

An obvious drawback of generic clustering algorithm lies in its instability that the variance of clustering results over repeated trails can be significant. To offset the uncertainty of clustering results, we evaluated the cluster stability through measuring pairwise cluster distance [5, 6]. The stability of a clustering algorithm with a certain choice of hyper-parameters $\boldsymbol{\theta}$ can be quantitatively assessed and therefore easily integrated as a loss function to assist the hyper-parameter optimization. Specifically, given only one data set $\mathcal{D}$, we considered a Hamming distance (see [5] for details) to compute the distance between two clusterings C and C':

$$d(C, C') = \min_{\pi} \frac{1}{N_{\mathcal{D}}} \sum_{\mathcal{D}} \mathbb{1}_{\{\pi(C(\mathcal{D}) \neq C'(\mathcal{D}))\}}, \qquad (1)$$

where $d(\cdot)$ denotes the distance function, $N_{\mathcal{D}}$ is the total number of voxels, $\mathbb{1}$ represents the Dirac delta function[7] that returns 1 when the inequality condition is satisfied and 0 otherwise, and function $\pi(\cdot)$ denotes the repeated permutations of data set $\mathcal{D}$ to guarantee the generalization of the stability measure [8]. Eventually, the expected distance between two clusterings over a pre-defined number of permutations on $\mathcal{D}$ was adopted as a stability score S, which was normalized to 0-1, with lower values indicating high stable clusterings. See Algorithm 1 for pseudo-code.

Algorithm 1: single-time stability measure

```
Divide D to 70% D_train and 30% D_test
for u iterations do
    fit the training cluster C^u_train with D_train
    fit the test cluster C^u_test with D_test
    predict with C^u_train, obtain C^u_train(D_test)
    predict with C^u_test, obtain C^u_test(D_test)
end
Compute S = (1/u^2) Σ_τ^u Σ_η^u d(C^τ_train(D_test), C^η_test(D_test))
```

2.1.2. Auto-encoder for state-space conversion

In addition to the stability analysis, an auto-encoder that enables the conversion from a dimensional space to the others was introduced to facilitate reliable clustering. An auto-encoder is an artificial neural network (ANN) that aims to learn a representation of a set of input data in an unsupervised manner [9]. A typical auto-encoder has three neural network layers:the input layer, hidden layer, and output layer [10]. Node number in the hidden layer N_{in} represents the new dimensionality after the conversion, and the number of nodes N_{out} are identical in both the input and output layers. We assume both N_{in} and N_{out} are part of the unknown variable set $\boldsymbol{\theta}$ to be optimized.

In this paper, we also included q, the quantile threshold that distinguishes outlier data points from the majority, as an unknown variable. Thus we have $\boldsymbol{\theta} = (N_c, N_{in}, N_{out}, q)$. Hyper-parameter tuning of $\boldsymbol{\theta}$ is non-trivial due to its enormous joint searching space. In practice, tuning of ML algorithms also requires machine learning experiences which could be infeasible for clinical experts. Nevertheless, it is highly likely that partial derivative continuous and gradient continuous may not exist in this type of hyper-parameter search space, which makes automatic tuning become a challenging task. Therefore, we introduced Bayesian optimization (BO), a sequential optimization technique based on Bayes' Theorem and Gaussian Processes (GP) to learn the non-parametric representation of the underlying black-box.

2.2. Bayesian optimization

BO aims to approximate the search space contour of $\boldsymbol{\theta}$ by casting the co-variance matrix of GP in light of data. It adopts an exploration-exploitation scheme to find the most probable candidate of $\boldsymbol{\theta}$ for surrogate function evaluation. As a result, the efficient search of sub-optimal hyper-parameter solutions become feasible under the BO framework. See [11] for details.

Formally, the proposed clustering process can be considered as a black-box system with input $\boldsymbol{\theta}$ and output S, where S is the stability score. Thus the training data of BO is $\mathcal{D}_B = \{\boldsymbol{\theta}_j, S_j\}_{j=1}^{J}$, where J is the number of data points and S can be computed by Equation (1). BO aims to minimize the (surrogate) function mapping $f : \mathcal{X} \rightarrow \mathcal{S}$, where $\mathcal{X}$ and $\mathcal{S}$ denote the input and output space respectively. We defined GP as: $f \sim \mathcal{GP}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$; where $\boldsymbol{\mu}$ is the $J \times 1$ mean function vector and $\boldsymbol{\Sigma}$ is a $J \times J$ co-variance matrix composed by the pre-defined kernel function $K(\cdot)$ over the data points $\{\boldsymbol{\theta}_j\}_{j=1}^{J}$.

A powerful GP algorithm requires carefully designed kernel function and its associated hyper-parameters. See [12] for an overview of GP and the kernel functions. In this paper, we adopted Matern 5/2 kernel to compose the co-variance matrix of f:

$$K_{M52}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2[1 + \sqrt{5r^2(\boldsymbol{\theta}, \boldsymbol{\theta}')} + \frac{5}{3}r^2(\boldsymbol{\theta}, \boldsymbol{\theta}')] \exp(-\sqrt{5r^2(\boldsymbol{\theta}, \boldsymbol{\theta}')}), \qquad (2)$$

where $r^2(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{\sigma^2}\sum_{m=1}^{N_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_m - \boldsymbol{\theta}'_m)^2$, σ_f denotes the signal standard deviation, $N_{\boldsymbol{\theta}}$ is the dimension of $\boldsymbol{\theta}$, and σ

describes the dimensional length scale.

BO introduced a so-called acquisition function $a(\cdot)$ to suggest the next $\boldsymbol{\theta}$ candidate to be evaluated by f. There are typically three types of acquisition functions: probability of improvement (PI), expected improvement (EI) and lower confidence bound (LCB) [11]. This paper employed the EI strategy, which searches for new candidates that maximizes the expected improvement over the current best sample. Suppose f' returns the best value so far, EI searches for a new $\boldsymbol{\theta}$ that maximizes the function: $g(\boldsymbol{\theta}) = \max\{0, f' - f(\boldsymbol{\theta})\}$. The EI acquisition function can then be expressed as a function of $\boldsymbol{\theta}$:

$$\begin{aligned} a_{EI}(\boldsymbol{\theta}) &= \mathbb{E}(g(\boldsymbol{\theta})|\mathcal{D}_B) \\ &= (f' - \boldsymbol{\mu})\Phi(f'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \boldsymbol{\Sigma}\mathcal{N}(f'|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal. To conclude, BO constructs a surrogate model described by f with an objective of maximizing the stability score S.

2.3. Clinical survival prediction

Once the N_c stable sub-regions were identified by clustering, we continued to characterize each sub-region and their global distribution using radiomic features. In order to extract k clinical features for the i_{th} patient, we measured K spatial radiomic features $F_i = (f_1, f_2, \ldots, f_K)$ based on N_c clusters.

In the previous study, radiomic features were extracted from the grey-level images[13], including the first-order statistics and second-order spatial distribution features. Particularly, the Haralick texture features obtained from the grey-level co-occurrence matrix (GLCM) are widely used. However, those features designed for the grey-level images may not be suitable to tumor sub-region analysis.

We instead proposed two types of feature, namely the multi-regional co-occurrence matrix (MRCM) and multi-regional run-length matrix (MRRLM). Specifically, MRCM summarizes the sub-regional volume and their co-existence pattern, while MRRLM describes the distribution pattern of multiple sub-regions that are different in size. In this paper, 10 features were extracted, including sub-region proportion, joint energy, entropy, an informational measure of correlation, categories diversity, short-run emphasis (SRE) and long-run emphasis (LRE), run-length non-uniformity, run variance and run entropy. Eventually, samples clustering algorithm, along with survival analysis, were applied to identify patient subgroups of higher-risk and lower-risk. Here we present the complete framework and corresponding algorithm for any glioblastoma cohort, as Algorithm 2 describes.

Algorithm 2: BO assisted clustering for survival prediction

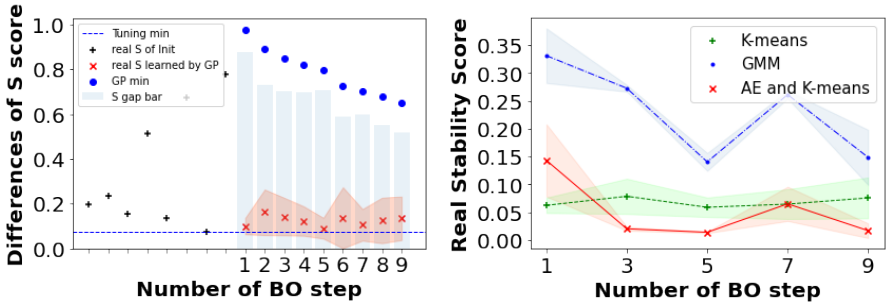
```
GP initialization;
while (BO did not converge)
or (not achieve satisfied stability) do
    Fit GP model with {θ_j, S_j}_{j=1}^J pairs
    Draw next θ by EI strategy
    Compute S by Algorithm1
    Estimate minimum of current GP surrogate model
end
Determine θ_best by well-trained GP model
Clustering D with θ_best to obtain N_c clusters.
Transfer D_i = (v, r, p, q) into D_i = (C_{N_c})
for each patient do
    Calculate MRCM and MRRLM matrices
    Extract textural features F_i = (f_1, f_2, ..., f_k)
end
Grouping patients into high-risk and low-risk groups
  with F_1, F_2, ..., F_N
Group survival analysis and clinical explanation
```

3. RESULTS

Three unsupervised learning algorithms were compared under the proposed framework, namely K-means, GMM and auto-encoder enhanced K-means (AE-K-means) with quantile threshold. Figure 1 (a) showed the stability performance of K-means clustering with hyper-parameters tuned by the optimal (minimum) point of GP surrogate model from BO. A lower stability score (SS) indicates a better clustering stability performance. As shown in Figure 1(a), the SS of initial points of which $\boldsymbol{\theta}$ is tuned manually scatter in a wide score range, while the ground-truth of $\boldsymbol{\theta}$ leaned by GP model fluctuates between small S score range as red crosses show. It is illustrated that the GP model could effectively learn the underlying black-box process to minimize SS with several initial points and a few BO steps. A clear decreasing trend of GP estimate minimum can be observed with BO steps, and the bar chart further indicated a decreasing gap between SS (in normalized form) of the GP model and that of the actual clustering process.

Thus, it is clear that with few step BO exploration, the GP surrogate model can alternatively determine the sub-optimal hyper-parameters for given the algorithm. Furthermore, we compared the ground-truth under hyper-parameters given by GP for the three algorithms in Figure 1 (c), proving the effectiveness of auto-encoder for facilitating reliable clustering via conversion from a dimensional space to the others.

Fig.2 shows that two patient subgroups were identified with distinct survival probability, which could validate the effectiveness of the proposed framework. Fig. 3. shows the sub-regions identified from two case examples, where different colors represent the different sub-regions identified from the patients. Intuitively, these sub-regions are highly overlapped with the regions of proliferating, necrotic, and edema tumor areas, respectively.



(a) S score gap between GP and truth (b) S score of clustering algorithms

Fig. 1. Figure (a) illustrates the difference of stability (S) score (in normalized form) between the global minimum point of the trained GP surrogate and the ground-truth minimum in the form of bar-chart using BO and K-means. The blue dots and red cross denote the S score of GP global minimum and the ground-truth values respectively. Besides, the black crosses scattered in the left are the initial points illustrate the manual parameters-tuning process. The blue dashed lines indicate the best S score can be achieved by manual fine tuning minimum, but contractually best minimum of GP global is zero (in normalized form). (b) demonstrate the S score performance under BO using three different unsupervised learning algorithms: K-means (green plus), GMM (blue dot), and AE+K-means (red cross).

Limitation: it is challenging to guarantee the BO optimized solution is the global optimal: (a) the GP needs to be converged first which requires a number of iterations, (b) even GP is converged, it is one possible representation of the clustering process, which may not be absolutely optimal.

Besides, based on the sub-region images we acquired, we deployed the methods in the framework (b) to extract devised clinical features and analyzing those features base on survival analysis. The results are shown in Fig. 4, the spatial feature(Group1) compartment showed a significant result(P = 0.0085) to distinguish patients into higher-risk and lower-risk subgroups.

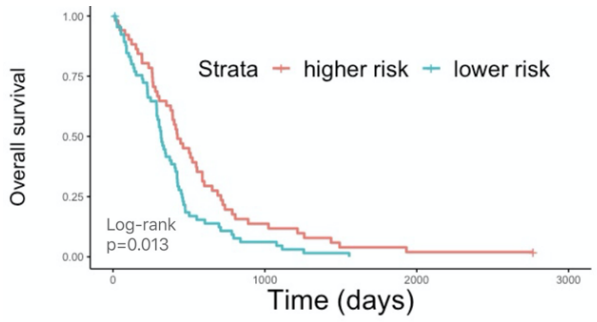
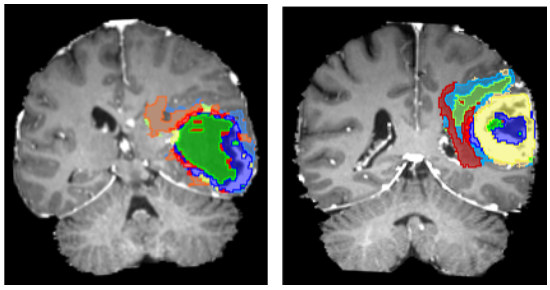


Fig. 2. Two patient subgroups were identified with distinct survival probability.



(a) higher risk (b) lower risk

Fig. 3. Two case examples with higher-risk and lower-risk respectively. Different colors represent the identified sub-regions. Two patients have distinct proportions of sub-regions,providing interpretable visualization for clinicians.

4. CONCLUSION

The paper is an interdisciplinary work that introduced an explainable framework to identify the stable intra-tumoral sub-regions while predicting patient survival. Bayesian optimization technique was applied to learn the black-box of the unsupervised learning process with the clustering stability score as the objective function. The global minimum of the trained Gaussian Process surrogate model is then used as the sub-optimal hyper-parameter solutions for reliable clustering. Based on the sub-regional MRI features, higher-risk and lower-risk patient subgroups can be derived, which could validate the utilities of the proposed framework and shows potential for future precision treatment.

5. REFERENCES

[1] Chao Li, Shuo Wang, Pan Liu, Turid Torheim, Natalie R Boonzaier, Bart RJ van Dijken, Carola-Bibiane Schönlieb, Florian Markowetz, and Stephen J Price, "Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals," *Neoplasia*, vol. 21, no. 5, pp. 442–449, 2019.

[2] E Sala, E Mema, Y Himoto, H Veeraraghavan, JD Brenton, A Snyder, B Weigelt, and HA Vargas, "Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging," *Clinical radiology*, vol. 72, no. 1, pp. 3–10, 2017.

[3] M Angulakshmi and GG Lakshmi Priya, "Brain tumour segmentation from mri using superpixels based spectral clustering," *Journal of King Saud University-Computer and Information Sciences*, 2018.

[4] Eva Patel and Dharmender Singh Kushwaha, "Clustering cloud workloads: K-means vs gaussian mixture model," *Procedia Computer Science*, vol. 171, pp. 158–167, 2020.

[5] Ulrike Von Luxburg, *Clustering stability: an overview*, Now Publishers Inc, 2010.

[6] Marina Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*, pp. 173–187. Springer, 2003.

[7] Lin Zhang, "Dirac delta function of matrix argument," *arXiv preprint arXiv:1607.02871*, 2016.

[8] Tilman Lange, Volker Roth, Mikio L Braun, and Joachim M Buhmann, "Stability-based validation of clustering solutions," *Neural computation*, vol. 16, no. 6, pp. 1299–1323, 2004.

[9] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] Mark A Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[11] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[12] Eric Brochu, Vlad M Cora, and Nando De Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *arXiv preprint arXiv:1012.2599*, 2010.

[13] Joonsang Lee, Rajan Jain, Kamal Khalil, Brent Griffith, Ryan Bosca, Ganesh Rao, and Arvind Rao, "Texture feature ratios from relative cbv maps of perfusion mri are associated with patient survival in glioblastoma," *American Journal of Neuroradiology*, vol. 37, no. 1, pp. 37–43, 2016.